**TABLE OF CONTENT**

**CHAPTER ONE: INTRODUCTION**

1.0 Introduction 1

1.1 Field and Subject Area of Study 1

1.2 Purpose of Research 1

1.3 Objectives of the study 2

1.4 Problem Statement 2

1.4.1 Prior Works and Study 3

1.4.2 Focus of the Study 5

1.5 Background and Justification of Study 5

1.6 Scope of the study 6

1.7 Research Methodology 6

1.8 Expected Outcome of the Research 7

1.9 Research Implementation Schedule 7

**CHAPTER TWO: LITERATURE REVIEW**

2.0 Introduction8

2.1 Definition and Description of Terms

2.1.1 Data Collection8

2.1.2 Data Preprocessing and cleaning 9

2.1.3 Features Encoding9

2.1.4 Features Scaling10

2.1.5 Dataset Splitting10

2.1.6 Regression10

2.1.7 Support Vector Machine10

2.1.8 Random Forest11

2.1.9 K-Nearest Neighbors (KNN)11

2.1.10 Artificial Neural Network(ANN)11

2.1.11 Linear Regression (LR)11

2.2 The Web Based Prediction System12

2.3 Empirical Review12

2.3.1 The Impact of Using Different Features on the Performance of the Regressions Models.14

2.3.2 Discussion14

2.3.3 Features Importance15

2.3.4 Experimental Settings15

2.4 Conclusion15

**CHAPTER THREE: RESEARCH METHODOLOGY**

3.0 Introduction117

3.1 Review Of Educational Data Mining Techniques Used In Profiling Srl

M19 impl19 test19 desi18 req18 def18 desidev18

3.2 Inclusion Criteria in21 su20 sya19

3.3 Systematic Review Results 22

3.3.1 Clustering Algorithms

3.3.2 Temporal Data Mining.

3.3.3 Other EDM Techniques

3.4 Sources Of Data And Feature Sets For Measuring Srlc ss23

3.5 Discussion on Systematic Review c23

**CHAPTER ONE**

**1.0 INTRODUCTION**

Education facilitates the learning or acquisition of knowledge, skills, values, morals, beliefs, habits, and self-development. Educational methods include education, training, storytelling, discussions, and directed questions. Education often takes place under the direction of educators; However, students can also train themselves. Education can take place in formal or informal settings, and any experience that has a formative effect on the way one thinks, feels, or acts can be considered educational. The teaching methodology is called pedagogy.

Formal education is generally formally divided into levels such as preschool or kindergarten, elementary school, high school, and then college, university, or education. In most regions, schooling is compulsory up to a certain age.

Education is a very important aspect of Ghana’s Government and it takes an expenditure of 3.99% of its GDP over 80perecent of its children population are enrolled in kindergarten and primary schools as cited by its ministry of education. The Ghana Education system is broken down into several levels kindergarten, Primary Secondary (Junior and High), and tertiary. Although there is a higher percentage of students enrolled in kindergarten to secondary than that of tertiary the percentage of students enrolled for tertiary education in Ghana is about 18.88% and just about 10% out of this 18.88% end up graduating. The purpose of this research is to predict student performance using their past records thereby giving insights on how to retain students and also increase the number of enrollments and graduation in tertiary education.

**1.1 FIELD AND SUBJECT AREA OF STUDY**

The Field of study for this research is Data Mining to be more Precise Educational Data Mining which is interdisciplinary of Computer Science and Statistics

**1.2 PURPOSE OF RESEARCH**

The purpose of this research is to develop a prediction system that can help predict students’ future performance using their past educational records

**1.3 OBJECTIVE OF STUDY**

**Main Objective**

The main objective of this research is to design and implement an academic performance prediction system

**Specific Objective**

* Create a web-based prediction system using educational data mining
* To allow this system to give future insight into students' academic performance
* Allow schools to make informed decisions from this prediction in order to retain and also improve the learning process of students

**1.4 PROBLEM STATEMENT**

As written in the introductory statement the percentage of students who graduate from secondary school and then enrolls in tertiary schools in Ghana is about 18.88% but not this percentage follow through and graduate. This is as a result of different factors ranging from financial problems, social problems, and most importantly to this research difficulty to pull through and understand and as a result dropping out of school.

The Main demographic targeted for this research is the tertiary level because of its low percentage enrollment ratio and its even lower graduation ratio.

This problem is prevalent in different countries in the world not just in Ghana, but advanced countries like the USA and China have started tackling this problem by using educational data mining to predict students' performance at the point of enrollment, during studies, and so on in order to make informed decisions and also assist the students to make an informed decision. The tertiary education system is a highly competitive system where average and less average students can get overwhelmed.

Predicting student performance can help predict weak students and help tertiary institutions management make strategies and decisions related to improving student performance. The institution management can take steps at an early stage to prevent failure or dropouts. Knowing the prediction, the weak student is also expected to improve his performance and get a better score.

**1.4.1 PRIOR WORK AND STUDY**

There are prior works in regards to predicting student academic performance but we will be taking an in-depth look at the work of Predicting Academic Performance in the School of Computing & Information Technology (SCIT) at the University of Technology Jamaica. There are different criteria used in screening students before giving them admissions and the UTJ uses in the School of Computing & Information Technology, the main admission criteria are appropriate scores in an aptitude test and pass in at least five Caribbean Examination Council subjects including Mathematics and English. The research carried out by the school of computing and information technology examined relationships between students' demographics, aptitude test scores, performance in first-year courses, their overall performance in the program, and so on.

The School of Computing and Information Technology (SCIT) at Jamaica University of Technology (UTJ) has had more than seven hundred (700) applicants per year for its programs over the past few years. Of that number, about two hundred and twenty-five are accepted. The faculty offers two courses, a Bachelor of Computing and Information Technology (BsCIT) and a Bachelor of Computing with Management Studies (BCMS). The BsCIT title is the more popular of the two and attracts around two-thirds of applicants. The selection of students for both programs includes the evaluation of the results of the tests of the Caribbean Examination Council

(CXC) and/or the General Certificate Ordinary Level (GCE O'Level) of the applicants, as well as the score obtained in the test of SCIT suitability. The current enrollment requirements for the programs are five (5) CXC or GCE-O`Level subjects, including mathematics and English, and preferably one natural science subject. With its application to admission ratio approximately 3:1 question was raised about the admission criteria for the program, are they selecting crème de la crème and to what extent does the admission decision variable affect students' academic performance, in addition, the failure rate in their Introduction to Programming (ITP) course has risen from 30% to just over 40%. This is a cause concern as ITP is a foundation course and a prerequisite for several other programming courses including Programming using C, Data Structure using C, and Advance Programming using Java. Although specific actions have been taken to address the course failure rates problems the question of whether or not it is possible to predict which student will perform well on the programs has persisted. And they asked questions like: If we can identify the factors that indicate which students are more likely to succeed in the programs then we will be able to optimize the selection process? They sampled 96 students, 68 males, and 28 females. To matriculate to the SCIT programs students are required to take an aptitude test and to have passed at least five subjects at CXC and/or GCE O’Level as earlier stated in their admission criteria. The subjects should include English Language, Mathematics and preferably a science subject. Fifty-five (57%) of the matriculated candidates passed between 7 and 10 O’Levels and 35 (approximately 37%) had between 4 and 6 subjects, just about six (6%) had between 1 and 3 subjects. the majority of the matriculated students did not have A 'Levels. Sixty-four (66.7%) of the students had no passes at A’Levels, while twelve (12.5%) had one subject pass and another twelve (12.5%) had two subject's passes. Regarding the GPA only four students (4.2%) received a first-class honor, 18 students or (approximately 19%) received upper second class, 48 students or (50%) received lower second-class honors and three students or (3%) received a pass. Twenty-three students (24%) discontinued. There were no failures. We examined a number of predictor variables; gender, scores in the aptitude test, passes in Chemistry, Physics, Math's, Additional Math's, GCE Advanced Level (A 'Level) Math, Accounting, Information Technology at both the O’Level and A’ Level, and the number of CXC, O’Level and A’ Level passes. We also included passes in first-year first-semester computing courses;

Introduction to Programming (ITP), and Computer Logics and Digital Design (CLDD). The dependent variable in this study was the student’s final year GPA. GPA was divided into five categories, first-class (3.45 – 4), upper second class (3.05 – 3.44), lower second class (2.40 – 3.04), pass (1.70 – 2.39), and fail (0 –1.69). All the data that was gathered was analyzed using stepwise multiple regression analysis. The stepwise approach eliminates variables already in the model that are no longer significant predictors. Three models were used to determine the prediction and each model was used with different predictors. The first model determined to what extent the aptitude test, and science subjects (Chemistry and Physics at O’Level and Physics at A 'Level related to academic performance. The results indicated that neither of these factors had any effects on GPA, as all were removed from the stepwise analysis. The second model examined the predictive value of Mathematics (both O’Level and A 'Level), Additional Mathematics, Principles of Accounting (both O and A-level), and Information Technology (both O and A levels) the result indicated that they were weak predictors for academic performance The third model examined the number of A-Level subjects passed and the number of O-Level subjected passed. The expectation is that people who have passed more O and A-level subjects would perform better than those who did not. This made them make informed decisions in the admission criteria and better improve how to classify the students in the two courses the school had to offer.

**1.4.2 FOCUS OF THE STUDY**

In an article written by the Univer newsroom on March 9, 2021, the president of Ghana Akufo Addo stated that the government's target was to increase enrollment for tertiary institutions from 18.88% to 40% by 2030 and this research aims at helping tertiary institutions make informed decisions about the admission process and several other things to improve students' performance thereby increasing its retention rate also.

This research tries to answer the following questions: "Can we predict the performance of students throughout their stay in tertiary institutions, using only their grades from secondary school without any socioeconomic factors. "and" can we identify freshman and second-year courses that are effective predictors for students' performance at the end of the degree?”. From an administrative point of view, it is easier to gather the marks of students than their socio-economic data. Therefore, if a reasonable prediction can be reached without socioeconomic data, it makes the implementation of the performance support system in tertiary institutions easier."

**1.5 BACKGROUND AND JUSTIFICATION OF STUDY**

The enormous growth of electronic educational data provides the opportunity to extract information that can be used to predict overall student success, predict student dropout rate, evaluate teacher and instructor performance, evaluate learning material according to the needs of students to improve and much more. This research aims to examine the latest trends in predicting student achievement in higher education and then apply and improve the methods used in mining these data. Higher education plays an integral role in the development of society (Pinheiro et al. 2015), so increasing academic success is a long-term goal of academic institutions.

Understanding and defining academic success is important for increasing student success rates. The definition of academic success is very complex and meaningful. Therefore, it is often misused in education and research. The study by York et al. (2015). which is often measured either through course evaluations or institutional surveys, persistence, which is measured by staying between certain academic years and graduation rates, acquisition of skills and competencies that can be measured by assignments, and course evaluations, Achievement of learning goals, which can also be measured by course assignments and evaluations, and finally Professional success, which can be determined by success rates, track record promotion, job satisfaction and occupations - goal achievement. A second key requirement for maximizing academic success is identifying the factors that impact academic achievement. Awareness of student success factors could help achieve the highest level of education (Yassein et al. 2017). Potentially, it can help provide a clear and meaningful description of the types of knowledge and behavior associated with proper performance. This awareness can be obtained by using data mining (DM) methods through educational materials. The practice of DM methods applied to educational data is known as Educational Data Mining (EDM) (Baker and Yacef 2009). It comes from a variety of fields including DM and machine learning, psycho-metrics, and other areas of statistics, information visualization, and computational modeling (Romero and Ventura 2007). In general, EDM refers to techniques and tools designed to automatically extract useful information and patterns from huge stores of data related to human learning activities in an educational setting (Nithya et al. 2016). These tools use machine learning algorithms, database systems, statistical analysis, and artificial intelligence. DM techniques include regression, clustering, class identification, association, and prediction.

The use of DM in education systems is presented as an iterative cycle of training, testing, and refinement of hypotheses, whereby systems can be adapted to support the specific needs of each participant in the educational process.

**1.6 SCOPE OF THE STUDY**

This project focuses on designing and developing an academic performance prediction system using Educational Data Mining. It will be limited to articles, documents, research, and journal papers only.

**1.7 RESEARCH METHODOLOGY**

For the purpose of this research, we will be analyzing two or more data mining techniques that are suitable for predicting academic performance and then choosing the best possible options. And will be mentioning the two most probable approach.

These techniques include

* The first approach is based on classification and regression algorithms that search for patterns in study-related data and also data about students' social behavior.
* The second approach is based on collaborative filtering techniques. We predict the final grades based on previous achievements of similar students

**1.8 EXPECTED OUTCOME OF RESULT**

This research aims at helping schools (tertiary institutions) predict student performance by using data they already have thereby giving them the power to make informed decisions that will better enhance student performance and increase the retention rate on students.

**1.9 THE RESEARCH IMPLEMENTATION SCHEDULE**

The research work would follow the timeline as provided below: -

* The first chapter would be developed in 2 weeks
* The second chapter will be developed in 2 weeks
* The third chapter will be developed in 3 weeks
* The fourth chapter would be developed in 3 weeks
* The fifth chapter will be developed in 2 weeks

**CHAPTER TWO**

**LITERATURE REVIEW**

**2.0 INTRODUCTION**

This chapter will be based on a review and discussion of relevant literature in connection with the research on Data Mining to be more Precise Educational Data Mining. I will carry out a censorious analysis of several components of the published body of Knowledge out through a summary, comparison of prior research studies, articles, findings, and observation of their limitation, advantages, and recommendation.

For the purpose of this study, the literature review is focused on functionalities, processes, and several components of the reviewed systems. The technology used and the functionalities provided by the system are the main focus when selecting the system for systematic review and comparison among the selected systems. The purpose of this literature review is to study the requirement specification, weaknesses, and strengths of selected systems. After considering these systems, ideas can be generated and implemented in the proposed system to overcome the weaknesses of this type of technology.

Besides this, functions that are commonly used among the selected systems could be selected as functions of the proposed system as well. The technology used by the selected systems is also studied to further enhance the proposed system.

**2.1 DEFINITION AND DESCRIPTION OF TERMS**

**2.1.1 Data Collection**

The dataset will be collected from the Accra Institute of Technology (AIT) and other systems. The data contains numeric data, which are the students’ course scores, and categorical data which are course names and the city. Each course has different columns depending on the course’s nature. Some columns were shared in all courses, such as the “mid-term” column. Other columns had been exceptional among courses, such as “participation”, “workshop”, and “discussion”. These columns had been blended in a single column called “mid activities”. Moreover, the dataset includes demographic facts approximately the scholars accumulated thru a questionnaire, which are:

Marital Status Data: Includes whether the student is married and has children. Health data: Includes whether the student is suffering from any medical condition. Residence data: Includes the city in which the student lives.

We have ten features; ten features are inputs and the last one is the output. The feature descriptions include the feature name, the description, and the domain. The domain means the numeric values of each feature. In features marital status, children, and health, “1” means “Yes” and “0” means “No”, while in level and class the domain of the feature is the range between specific numbers.

**2.1.2 Data Pre-processing and Data Cleaning**

Pre-processing is an important step in data mining. The purpose of data preprocessing is to convert the data into a suitable form that can be used by algorithms. Three main pre-processing steps have been applied to the dataset which are data cleaning, features encoding, and features scaling. The pre-processing was implemented using Python language and Microsoft Excel.

Data in the real world is often noisy and unstructured. The data cleaning step aims to fix anomalies in the data by filling in missing values and smoothing out noise when finding outliers. Handling Missing Values: In our dataset, several students did not fill out the questionnaire which led to missing values in the demographic data columns.

**2.1.3 Features Encoding**

In Machine Learning models, all inputs and outputs are required to be numerical variables. Therefore, when there is categorical data, it must be encoded before using it in the model, and this is called features encoding. So, encoding is the pre-processing of the categorical data when working on a model of ML algorithms. There are several techniques to encode categorical.

features, such as Label Encoding, One-Hot Encoding, Frequency Encoding, Ordinal Encoding, Binary Encoding, Hash Encoding, and Mean/Target Encoding.

In our dataset, we have two categorical features, namely “course” and “city”. The course feature contains the abbreviation for the course names, and the city feature is the name of the city where the student currently lives. To encode these two features into numeric features, we started by applying Label Encoder and One-Hot Encoding techniques to determine which technique suits our dataset. Label Encoder gave us the lower MAPE, so we used it in our dataset.

**2.1.4 Features Scaling**

In our dataset, we have two categorical features, namely “course” and “city”. The course feature contains the abbreviation for the course names, and the city feature is the name of the city where the student currently lives. To encode these two features into numeric features, we started by applying Label Encoder and One-Hot Encoding techniques to determine which technique suits our dataset. Label Encoder gave us the lower MAPE, so we used it in our dataset.

**2.1.5 Dataset Splitting**

The dataset is split into training datasets and testing datasets. The training dataset is used to build the model and the testing dataset is used to evaluate the model. The model has been validated by using different validation techniques which are: percentage split and cross-validation. The validation technique that gave better results in terms of prediction’s MAPE has been chosen.

**2.1.6 Regression**

Regression techniques are used to predict continuous results rather than predicting individual class labels. The purpose of this study is to predict the student's overall score. Therefore, the prediction problem is considered a regression problem. Compare different algorithms (that is, SVM, RF, KNN, ANN, and LR) to determine the appropriate algorithm for predicting the total student score for each course. Regression model-based predictions take features as input, run processes on them, and predict total scores as output. We then used metrics to compare the performance of all regression models. The algorithm with the lowest predicted MAPE is used in the prediction system.

**2.1.7 Support Vector Machines (SVM)**

SVMs are one of the most popular supervised ML algorithms used primarily for classification tasks (Bithari, T.B 2020). It was developed by Vapnick in 1995 to deal with the problem of prediction and pattern recognition and for analyzing and mapping both linear and nonlinear functions.

The basic idea of how the SVM algorithm works is to find the hyperplane that can separate data belonging to two classes with a maximum margin, and builds a hyperplane, or a group of hyperplanes (classes) in a high dimensional space. It classifies the object into categories, above or lower plane depending on the features of the object and by using kernel techniques, it can convert nonlinear to linear before partitioning (Adejo O.W 2018).

**2.1.8 Random Forest (RF)**

RF is a supervised learning algorithm that was developed by Leo Breiman in 2001. RF is a form of ML method based on the aggregation of many decision trees used for regression and classification. In general, the more trees in the forest, the more robust the prediction.

RF has become a standard tool for data analysis, making it an effective tool for prediction (Akar, Ö 2012). Decision tree regression predicts a number as the output of specific data. RF regression receives the average of these predictions as the "final" output.

**2.1.9 K-Nearest Neighbors (KNN)**

The KNN method was first explained in the early 1950s. KNN is a supervised ML algorithm that can be used to solve problems such as classification and regression. KNN is purely lazy and maintains the entire training set. Ferring all inductive generalization attempts before regression time (Han, J. 2011).

**2.1.10 Artificial Neural Network (ANN)**

McCulloch's work on simulating the biological nervous system culminated in the development of the ANN method in the 1940s. An NN is a set of interconnected units whose properties are determined by the topology of the network and the properties of the neurons. NN is one of the most widely used and efficient learning systems today. ANN learns, trains, and changes in the same way that humans do in the brain.

**2.1.11 Linear Regression (LR)**

Regression methods are used to describe the relationship between response and explanatory variables. LR is one of the simplest ML algorithms for supervised learning techniques and is used to solve regression problems.

LR is used to predict continuous dependent variables using independent variables. When predicting using one independent variable, it is called a single LR, and when there are three or more independent variables, it is called multiple LRs.

**2.2 THE WEB BASED PREDICTION SYSTEM**

The main goal of developing a student academic performance prediction system is to enable decision makers and university advisors to easily and quickly predict a student's overall score. After training the model using various algorithms: SVM, RF, KNN, ANN, and LR. The RF algorithm has the lowest percentage of MAPE and was chosen as the basic algorithm for website prediction systems.

This system allows users to enter student scores and demographics and use predictive models to predict the student's overall score. The table will then display a report containing all the information about the student (that is, general and demographic information, expected total score, expected grades). You can also print this report.

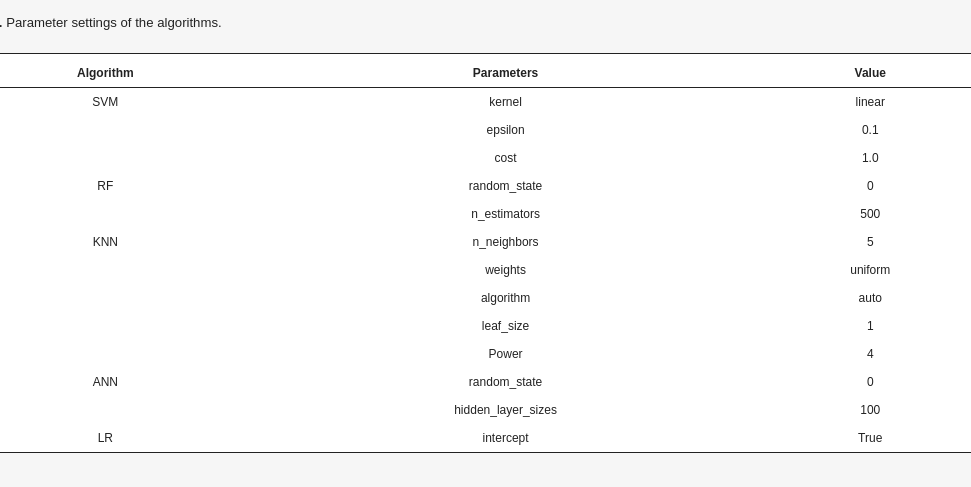
At the start of the system flow, we need a condition to call two predictive models and handle missing value errors.

The system operates under the following conditions:

1. If the values of the mid activities and lab scores are null, then the first prediction model is called, which is based on the midterm exam score only from academic data. Then it is transferred to the report template.
2. If the value of the mid activities score is null and the value of the lab score is not null, the user is warned that they must be entered together.
3. If the value of the mid activities score is not null and the value of the lab is null, the user is warned that they must be entered together.
4. If all the data are entered, then the second prediction model is called based on all three academic data. Then it is transferred to the report template.

**2.3 EMPIRICAL REVIEW**

Machine Learning models were developed using five algorithms, namely SVM, RF, KNN, ANN, and LR. In the experiments, Adjusting the hyper-parameters of each algorithm is out of the scope.



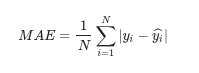
Performance Metrics

The following performance metrics were used to determine whether the model is accurate or not.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is one of the most common performance metrics. It is used to calculate the prediction error of the model. The MAE measures the average magnitude of the errors in a set of predictions.

MAE is given by the following Equation (1):



where yi represents the actual value, yiˆ represents the predicted value of yi, and N represents the number of instances.

Mean Absolute Percentage Error (MAPE)

The MAPE is calculated as follow.



where MAE represents the value of MAE and yi represents the actual value.

**2.3.1 The Impact of Using Different Features on the Performance of the Regressions Models**

The objective of this research is to know the best features that affect students’ academic performance, through which it is possible to predict the total score of students early, and take necessary measures when failures are discovered. Hence, two scenarios in which to investigate the effect of the features set on model performance were set. The first scenario when using the first dataset (DS1), which contains only 8 features, the midterm exam score and all demographic features (i.e., course, level, mid-term, marital status, children, health situation, living place, and class which is the total score for students). The second scenario when using the second dataset (DS2), which contains all the academic and demographic features (i.e., course, level, mid-term, mid activities, lab, marital status, children, health situation, living place, and class which is the total score for students).

**Scenario 1: Midterm Exam Score Only and All Demographic Features**

In the first scenario, predict the total score by only one of the academic features, which is the midterm exam score, and all demographic features.

**Scenario 2: All-Academic and Demographic Features**

The second scenario predicts the total score by using three academic features: midterm exam, mid activities, lab scores, and all demographic features.

**2.3.2 Discussion**

Two scenarios were conducted to find out the effect of a set of features on the performance of the regression's models. It can be noticed from Scenario 1 and Scenario 2 that the performance of the regression models that are using the DS2 achieve lower MAPE scores. The reason for this is the increase in the number of important features used for prediction.

**2.3.3 Features Importance.**

The dataset may contain many features, but some may not affect students’ academic performance. Hence, it is recommended to know the most important features. The purpose of this is to identify the features that have a significant impact on students’ academic performance and through which it is possible to discover whether the student will fail. After training the prediction model, the RF calculated the feature's importance by ranking the importance of the features. Features ranking aims to sort the features regarding the index’s quality that reflects the individual relevance of a feature [**31**]. Ranking techniques were used to rank features, and it assigned rank values to each feature according to its influence.

**2.3.4 Experimental Settings**

Machine Learning models were developed using five algorithms, namely SVM, RF, KNN, ANN, and LR. In the experiments, Adjusting the hyperparameters of each algorithm is out of the scope.

**2.4 CONCLUSION**

This project aims to develop a Machine Learning model to predict the students’ academic performance at the early stages based on the course level. The students’ data were collected from the Accra Institute of Technology (AIT) containing academic and demographic factors. Several algorithms (SVM, RF, KNN, ANN, and LR) were applied to predict students’ total scores. The RF algorithm obtained the lowest MAPE and then it was adopted in the Web-Based prediction system. Academic factors had a higher impact on students’ academic performance. Other features such as marital status, children, health, and the living place had little effect on students’ academic performance.

The model can be used as an early warning mechanism to identify and improve the performance of at-risk students. With the help of Machine Learning methods, tutors are in a position to know which of their students will complete a course with the lowest error rate and which student should be admitted into the school. In addition, creating a web page helps tutors to enter data easily, and print a report with academic students’ performance. Moreover, this paper contributed to considering a new factor that has not been studied in previous studies, which is living place (city). In future work, the system can be developed to add more features such as the academic history of the students, the effect of the English Language on the students’ academic performance, and frequent absence.

**CHAPTER THREE**

**RESEARCH METHODOLOGY**

**3.0 INTRODUCTION**

This chapter takes into consideration the various findings made in the review of the existing system. Methodology is the systematic, theoretical analysis of the methods applied to a field of study. It comprises a collection of procedures, techniques, tools and documentation aid, which will help system development in an effort to implement a new system. System analysis and design will be performed to determine if the proposed system is feasible to design and implement and will meet user requirements in ways that eliminate the weakness of the existing system.

This project aims to build a Machine Learning model to research Data Mining to be more Precise Educational Data Mining (EDM) which is interdisciplinary of Computer Science and Statistics with specified requirements using python libraries. This involved researching the Data Collection, Data Pre-processing and cleaning, and the use of different Algorithm models such as the SVM, RF, KNN, ANN, and LR. The chosen approach is to integrate

all the necessary modules to achieve a result of proper prediction of Student performances, (written in python) and uses of various machine learning libraries.

To address the research questions in the current study, we used a mixed-method approach. First, a systematic review of the literature on current EDM techniques used to profile SRL was carried out. The review followed five steps of the systematic review methodology (Khan et al., 2003). The review stages included (a) framing the research questions, (b) identifying relevant literature, (c) setting the articles’ assessment criteria, (d) presenting review results, and (e) discussing the results. This review formed the foundation for the second study which involved an experimental evaluation of EDM algorithms to establish the optimal algorithm to identify SRL profiles from a dataset obtained from the Open University in the UK.

Finally, correlation analysis was used to identify the association between the SRL profiles and students’ academic performance.

The articles reviewed in this study were repeatedly searched in international journals and

databases, including Google Scholar, SCOPUS, Science Direct, Elsevier, ERIC, IEEE

Xplore, and ACM Digital Libraries. We searched for articles using keywords: "educational data mining techniques" and "learner analysis" and "measurement of self-adjusting learning" and "evaluation of self-adjusting learning" and "clickstream data" and "student behavior". And "online learning" and "self-regulation". A total of 742 papers have been identified. Read the full text of each article, Khan et al., 48 articles have been deleted after applying the inclusion criteria described in 2003. Twenty-four articles, twelve journal articles, and twelve conference articles met the selection criteria.

**3.1 DESIGN AND DEVELOPMENT PROCESS OF ACADEMIC PREDICTION SYSTEM**

To design the prediction system, we looked into current thoughts and research on the best possible development process approach, and these approaches are listed below

**3.1.1 Defining and Understanding the Problem.**

A problem statement is a brief description of the problem this project is trying to address. A problem statement identifies the current state, the desired future state, and the gap between the two. A problem statement is an important communication tool that helps ensure that everyone involved in a project understands the issues to be addressed and why this project is important. And the problem definition has been stated clearly in chapter 1 of this research paper.

**3.1.2 Requirement Analysis of the entire System.**

Once the problem definition and understanding are complete, the next step is to define and document the requirements for the way the system will work. This is done through an SRS (Software Requirements Specification) document that contains all the product requirements that need to be designed and developed during the life cycle of the project.

**3.1.3 Design and Development of the System.**

In this Phase, the SRS document created is transformed into a more logical structure so that it can later be implemented in a programming language. Operations, training, and maintenance plans are all created so that when developing we know what to do at each subsequent stage of the cycle. The development phase is where the actual codes are written and the application is built according to previous design documents and outlined specifications. Different programming language are usually used to develop a system for this system we will be using python to achieve the machine learning part and HTML, CSS and JavaScript to achieve the frontend.

**3.1.4 Testing and Evaluation**

Evaluation includes performing functional testing such as unit testing, code quality testing, integration testing, system testing, security testing, performance testing, acceptance testing, and non-functional testing. If an error is detected, the developer will be notified. A verified (actual) error is fixed and a new version of the software is created. The best way to ensure that all tests run regularly is to implement automated tests. Continuous integration tools support this need.

**3.1.5 Implementation and Integration.**

After testing, the overall design of the software is put together. Various modules or designs are integrated into the main source code through the efforts of developers. A training environment is typically used to detect further errors and defects. Information systems are integrated into the environment and eventually installed. After this phase, the software is theoretically ready for the market and available to all end users.

**3.1.6 Maintenance.**

In the maintenance mode we practice the necessary activities to resolve issues reported by end users. Additionally, the developer is responsible for implementing any necessary changes to the software after deployment. This may include addressing remaining bugs that could not be patched prior to release, or resolving new issues that surfaced based on user reports (Synops 2015). Larger systems may require longer maintenance windows than smaller systems.

**3.2 System Analysis Materials to Be Used**

(Spence and Grout, 1978) proposed that systems analysis and design (SAD) should be part of the computer science curriculum. Validate requirements using a system modeling and prototyping approach. Define requirements as a basis for why we should develop an academic prediction system, and use survey and interview techniques as information gathering methods in understanding the scope of our investigation.

**3.2.1 Survey as method of data collection.**

A survey is a simple tool for gathering information. A poll usually consists of a series of questions designed to assess a participant's preferences, attitudes, traits, and opinions about a particular topic. As a research method, questionnaires allow concepts to be counted or quantified. That is, use a wider sample or subset of the audience, which allows the findings to be applied to a wider population. For example, in a given year he has 100,000 unique users of his website. If he collects information from 2,000 of these users, he can confidently apply that information to all 100,000(Chris Gray 2016). There are different types of questions in surveys, depending on the circumstances of the survey and the type of information you want to access. Many surveys combine open-ended and closed-ended questions, such as rating scales and semantic scales. This means it can be used for qualitative and quantitative research. Surveys come in two main forms. Paper form or online form. Paper surveys are a more traditional method of data collection and can easily lose data. Paper forms are also cumbersome to organize and process.

When it comes to the digital space, surveys can be used for a variety of purposes, including:

* Collect feedback during a live product or pilot.
* Research because people visit her website and evaluate their experience of that visit (e.g. True Intent surveys).
* Quantify the results of qualitative research activities such as contextual research and interviews. When Evaluating usability such as System usability scale.

Surveys can be effective for identifying:

* Who your users are;
* What your users want;
* What they purchase;
* Where they shop;
* What they own

**Benefits of Surveys are:**

* It provides information to better understand end-users in order to develop better products.
* Reduce the risk of designing the wrong or inappropriate solution for your users.
* Provides confidence to stakeholders that the design is or will be effective. Collecting a larger sample size compared to qualitative research often speaks the language of business people. Like it or not, research often gives the impression that more is better.

**Disadvantages of Surveys are:**

* The validity of the research data can be affected by survey response bias.
* High survey dropout rates can also affect the number of responses received in your survey.

**3.2.2 Interview as a method of data Collection**

Interviews are used to collect data from small groups on a wide range of topics. You can use structured or unstructured interviews. Structured interviews are similar to surveys, with the same questions asked in the same order for each topic and multiple-choice answers. In unstructured interviews, questions vary by topic, may depend on answers to previous questions, and do not have fixed answer options. Where face-to-face interviews cannot be conducted, study participants/respondents may be interviewed using video conferencing tools in addition to regular telephone calls. You can safely use the different video conferencing tools such as: Microsoft Teams, Google meet etc.

The main purpose of interviews as a data collection tool is to collect data comprehensively and centrally. As Pauline Young pointed out, the purpose of the interviews could be the exchange of ideas and experiences, the collection of information related to a very wide range of data in which respondents repeat the past and define their present and future desires. there is. discuss options

**Importance of Interview:**

* It is the most suitable method for assessing personal qualities.
* It has clear value for diagnosing and treating emotional problems.
* This is one of the essential foundations for implementing the recommended procedures.
* It provides information that complements other data collection methods.
* It can be used not only for observation but also for verifying information obtained by the response method.

**Types of Interviews**

There are different types of interviews used to collect survey data. Interviews are either structured or unstructured, depending on whether a formal questionnaire is formulated, and questions are asked in a pre-determined order. Interviews can be direct or indirect, depending on whether the purpose of the question is explicitly stated or deliberately hidden. Classifying these two traits against each other yields four different types of interviews. That is, interviews can be (1) structured direct, (2) unstructured direct, (3) structured indirect, or (4) structured It may be an indirect one. Types (1) and (2) are basically objective types. (3) and (4) are subjective types.

**3.3** **System Design and Implementation Methodology**

The Adopted methodology for the proposed system is Agile. Agile software development refers to a set of iterative software development methodologies in which requirements and solutions evolve through collaboration between self-organizing cross-functional teams (Jack Mchill 2005). Agile methods and processes generally promote a disciplined project management process that encourages frequent inspection and adaptation, a leadership philosophy that encourages teamwork, self-organization, and accountability, a set of engineering best practices designed to allow for the rapid delivery of high-quality software, and a business approach that aligns development with customer needs and company goals. Agile Methodology refers to a practice that encourages continuous iteration of development and testing throughout the project's software development lifecycle. Unlike the Waterfall model, both development and testing activities are carried out concurrently in the Agile model of software testing. The Agile software development methodology is one of the most straightforward and efficient methods for translating a vision for a business need into software solutions. Agile software development approaches use continuous planning, learning, improvement, team collaboration, evolutionary development, and early delivery. It promotes adaptable responses to change.

Four core values are emphasized in agile software development.

* Interactions between individuals and teams regarding processes and tools
* Working software trumps extensive documentation.
* Customer involvement in contract negotiations
* Responding to change by sticking to a plan

**Advantages of Agile Methodology**

* It is flexible and adaptable
* Lower Cost
* Improved Quality
* End-user satisfaction

**Disadvantages of Agile Methodology**

* Scalability
* Skill Required

**3.4** **System Security**

The problem of system security is divided into four interconnected issues: security, integrity, privacy, and confidentiality. As stated in the data rationale, they determine the file structure, data structure, and access procedures.

1. System security refers to the technical enhancements and procedures implemented in hardware and operating systems to protect against intentional or unintentional damage from a defined threat. Data security, on the other hand, is the protection of data from loss, disclosure, modification, and destruction. The application selects the device number (IMEI number) and sends it to the API database.

2. System integrity refers to the proper operation of programs, as well as appropriate physical security and protection from external threats such as eavesdropping and wiretapping. Data integrity, on the other hand, ensures that do not differ from others, and how the organization can be protected from unwanted, unfair, or excessive dissemination of information about it.

3.The term confidentiality refers to the special status given to sensitive information in a database in order to minimize potential invasions of privacy. It is a characteristic of information that defines its need for protection. The technical means of providing such protection is system security. In contrast, privacy is primarily a matter of how information is used.

**3.5 CONCLUSION**

In conclusion, methodology is the systematic, theoretical analysis of the methods applied in developing a software. Agile methodology has been chosen as the adopted methodology for this project because of the nature of the project and the advantages attached to Agile methodology with respect to the project. Various information gathering and system analysis and design techniques are considered so as to come up with an accurate system.